Week 2 - Part 2 : What is statistics?

pending tasks	
type	

What is statistics?

def statistics - (t = 2:49)

Statistics is required for collecting, processing, describing and modelling data. The purpose of modelling is to draw inferences about data (in the context of statistics). Statistics is used to study a large collection of people or objects. To draw inferences from a large group, a smaller representative subset is surveyed.

Key terms and definitions:

def population - (t = 8:47)

def sample - (t = 9:01)

def parameter - (t = 9:50)

def statistic - (t = 10:12)

Population : Total collection of all objects of interest.

Sample : Subgroup of the population used to draw inferences about the population.

Parameter : The quantity estimated from the whole population.

Statistic : It is the quantity estimated from a small sample.

Fig.1 Parameter vs statistic.

How to select a sample?

The sample selected should be a good representative of the population.

Different methods of sampling are:

- Simple random sampling.
- Stratified sampling.
- Cluster sampling.

A sampling strategy is said to be truly unbiased if every element in the population has an equal chance of being a part of the sample.

How to design an experiment?

While studying the effect of one variable on another, the effect of lurking variables must be nullified.

How to describe and summarize data?

Charts and plots can be used to describe the general trends in data along with the summary statistics of the data.

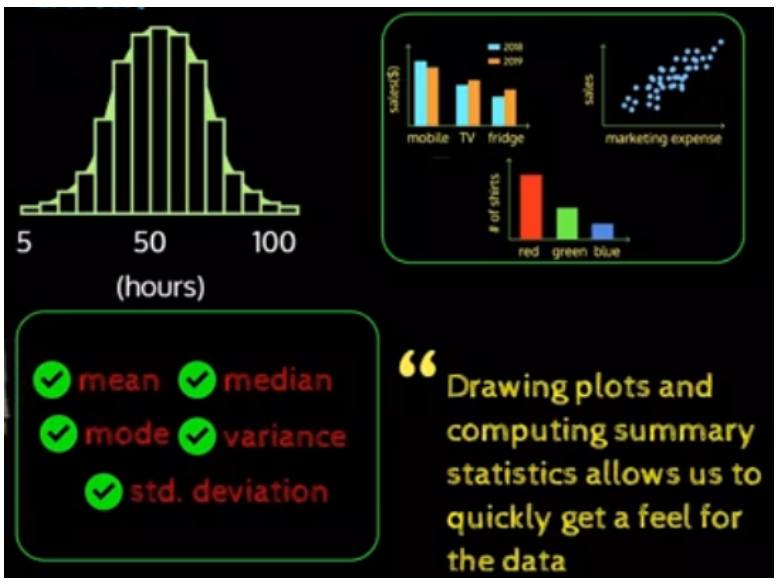

Fig.2 Describing data and summary statistics.

The branch of descriptive statistics deals with describing and summarizing data.

Why do we need probability theory?

def probability theory - (t = 1:47)

It is required to find the chance that the trend observed in sample data is present in the population.

How do we give guarantees for estimates made from a sample?

def point estimate - (t = 4:08)

def interval estimate - (t = 4:14)

This is given by calculating the point estimate and interval estimates and distribution sampling statistics.

What is a hypothesis and how do we test it?

It is a conclusion drawn from the statistic observed. Using this robust statements have to be made in regard to the entire population.

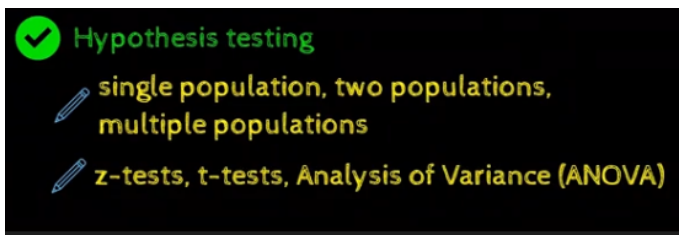

Fig.3 Hypothesis testing.

How to model relationships between variables?

This is done using statistical modelling. It assumes a very simple relationship between variables. Conclusions on the errors in estimate have to be drawn.

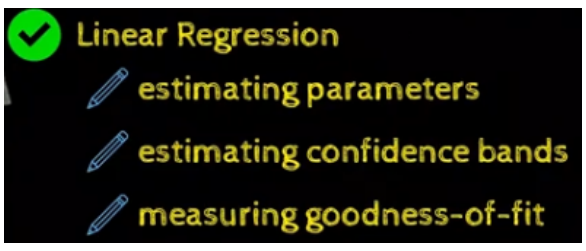

Fig.4 Example of model in the course.

How well does the model fit the data?

This is to estimate the accuracy of the model for given data.

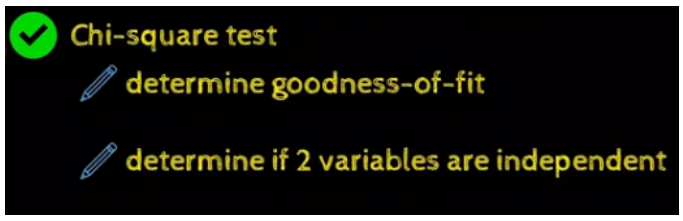

Fig.5 Estimating goodness of a model.

Summary : Week2 - Part2

The following topics would be explored in the course.

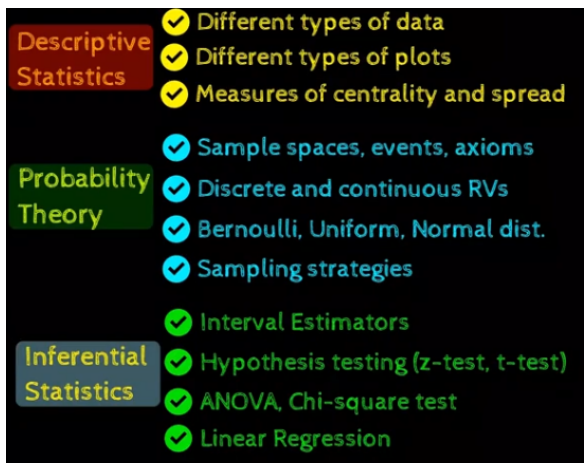

Fig.6 Topics to be covered.

- Statistic is the quantity estimated for a sample while the parameter is for the population.
- Various sampling methods are used to select a sample from a population.
- Drawing charts, plots and calculating summary statistics helps in quick understanding of the data.
- To make robust statistical inferences from data, probability theory is used. This aids in providing guarantees for estimates and in hypothesis testing.
- Statistics can also be used to understand the relationships between various variables.

MCQ

1. Why is adoption of machine learning models in medical environments slow?
    1. It is a very sensitive domain.
    2. It is generally a tedious process to acquire relevant dataset.
    3. It is difficult to establish accountability.
    4. **All of the above**
2. Data visualisation and summary statistics help in
    1. Understanding the general trend in the data.
    2. Understanding the spread of data.
    3. Understanding the relationship between variables.
    4. **All of the above.**
3. A statistic is
    1. a population characteristic
    2. **a sample characteristic**
    3. unknown
    4. normally distributed
4. A parameter is

1. **a population characteristic.**
2. a sample characteristic
3. unknown
4. normally distributed

This scenario applies to Questions 5 to 7: A randomized experiment was done by randomly assigning each participant either to walk for half an hour three times a week or to sit quietly reading a book for half an hour three times a week. At the end of a year the change in participants' blood pressure over the year was measured, and the change was compared for the two groups.

5. This is a randomized experiment rather than an observational study because:

1. Blood pressure was measured at the beginning and end of the study.
2. The two groups were compared at the end of the study.
3. **The participants were randomly assigned to either walk or read, rather than choosing their own activity.**
4. A random sample of participants was used.

6. The two treatments in this study were:

1. **Walking for half an hour three times a week and reading a book for half an hour three times a week.**
2. Having blood pressure measured at the beginning of the study and having blood pressure measured at the end of the study.
3. Walking or reading a book for half an hour three times a week and having blood pressure measured.
4. Walking or reading a book for half an hour three times a week and doing nothing.

7. If a statistically significant difference in blood pressure change at the end of a year for the two activities was found, then:

1. It cannot be concluded that the difference in activity caused a difference in the change in blood pressure because in the course of a year there are lots of possible confounding variables.
2. Whether or not the difference was caused by the difference in activity depends on what else the participants did during the year.
3. It cannot be concluded that the difference in activity caused a difference in the change in blood pressure because it might be the opposite, that people with high blood pressure were more likely to read a book than to walk.
4. **It can be concluded that the difference in activity caused a difference in the change in blood pressure because of the way the study was done.**

8. Which of the following is/are true?

1. **A sample statistic may change every time it is measured, but a population parameter remains the same.**
2. A population parameter changes every time it is measured but a sample statistic remains fixed across samples.
3. **Sample statistic is measured as calculation of population parameter may be infeasible due to large sample size.**